DOCUMENT RESUME

ED 079 325 TM 002 923

AUTHOR Gray, William M.
TITLE Development of a Piagetian-Based Written Test: A Criterion-Referenced Approach.
PUB DATE Feb 73
NOTE 23p.; Paper presented at Annual Meeting of American Educational Research Association (New Orleans, Louisiana, February 25-March 1, 1973)

EDRS PRICE MF-$0.65 HC-$3.29
DESCRIPTORS Age Differences; Cognitive Development; *Cognitive Tests; *Criterion Referenced Tests; *Measurement Instruments; *Response Style (Tests); Speeches; Statistical Analysis; *Test Construction
IDENTIFIERS *Piaget (Jean

ABSTRACT

An attempt was made to develop and validate a Piagetian-based written test with successful use of the logic of specific Piagetian tasks defined as the criterion. Ninety-six randomly selected nine to sixteen year olds, stratified by age, were individually presented the Piagetian tasks of pendulum, balance, and combinations, and group administered a 36-item logically equivalent written test. Restuls indicated that a Piagetian-based written test was successfully constructed. Discussion focused on future lines of research and the possible uses of such a test. (Author)

ED 079325

U.S. DEPARTMENT OF HEALTH,
EDUCATION & WELFARE
NATIONAL INSTITUTE OF
EDUCATION
THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

DEVELOPMENT OF A PIAGETIAN-BASED WRITTEN TEST: A CRITERION-REFERENCED APPROACH

William M. Gray

University of Dayton

TM 002 823

Paper presented at the annual meeting of the American Educational Research Association, New Orleans, February 1973.

FILMED FROM BEST AVAILABLE COPY

DEVELOPMENT OF A PIAGETIAN-BASED WRITTEN TEST:

A CRITERION-REFERENCED APPROACH

An attempt was made to develop and validate a Piagetian-based written test with successful use of the logic of specific Piagetian tasks defined as the criterion. Ninety-six randomly selected 9 - 16 year olds, stratified by age, were individually presented the Piagetian tasks of pendulum, balance, and combinations and group administered a 36-item logically equivalent written test. Results indicated that a criterion-referenced approach to constructing a Piagetian-based written test of cognitive development is possible and that the average age of change from concrete to formal operations is consistent with previous research.

DEVELOPMENT OF A PIAGETIAN-BASED WRITTEN TEST: A CRITERION-REFERENCED APPROACH[1]

Traditionally, assessment of cognitive constructs has been based on the work of Binet, with two methodological approaches: individual or group-administered tests. These approaches have been based on psychometric rigor and convenience, with little regard to understanding why a subject performed as he did. An individual's assessment and subsequent rating has been dependent on the mastery of specific information and on his position relative to a norm group within the normal curve model of probability. Consequently, if an individual did not know that the Koran is the Islamic holy book, or that the Apocrypha were the disputed books in the Bible, he did not receive any credit toward a rating of his cognitive prowess for those items. Because such tests generally have not been based on a theory of psychological development, they have not been adequate in assessing the development of specific constructs and, in reality, have caused many problems of interpretation, especially within the school situation.

Piaget has used a variation of the individual testing situation (his methodé clinique) and has attempted to assess cognitive constructs which do not depend upon knowing specific elements of knowledge or upon how an individual performs relative to a norm group within the normal curve model; rather, his work has focused on assessing cognitive constructs that are necessary

for competent interaction with the world, generally not teachable, and develop in all individuals at different rates. Although cognitive construct development is continuous, there are durations of time (periods) within which the individual's cognitive behavior is fairly stable and qualitatively different from the behavior of the other periods. Within each period of stability, Piaget distinguishes two subperiods: a beginning subperiod, where the individual begins to manifest the logical cognitive characteristics describing the overall period, but fails to consistently manifest those characteristics and consequently at times regresses cognitively and manifests characteristics of an earlier period; and a second subperiod, where an individual consistently manifests the logical cognitive characteristics of the overall period, generally does not regress cognitively, and manifests sporadically the logical cognitive constructs of the first subperiod of the next period (Inhelder & Piaget, 1958). Although the logical cognitive structures of each subperiod of an overall period are similar, they are also different, and, as such, enable an individual to solve different logical problems at different periods in his life.

Unfortunately, Piaget's methodé clinique, like the individual method within the Binet tradition, is very time consuming and difficult to employ. Much information can be obtained about one person per unit of time, but very little information can be obtained about many people in the same unit of time. A Piagetian-based group-administered written test of logical cognitive devel-

opment would be able to provide much information about many individuals per unit of time. Such a test would be a criterion-referenced one, as it would provide " . . . scores that tell what kinds of behavior individuals with those scores can demonstrate [Nitko, 1970, p. 38]." The test could be designed with several scales, each scale corresponding to the development of an overall specific logical cognitive behavior, while subscales within a scale would correspond to the developmental logical behaviors associated with specific periods and subperiods.

The present study was an attempt to construct a group-administered written test that would assess the same developmental logical constructs as those assessed by specific Piagetian individually-administered tasks by comparing response "patterns" on the written test with response "patterns" on the Piagetian tasks.

Method of Investigation

Construction of Written Instrument

Three scales, each corresponding to a specific set of developmental logical cognitive behaviors, comprised the test. Each scale was constructed with the suggestions implied by Glaser and Klaus (1962) and Glaser and Cox (1962) for criterion-referenced measures and the recommended specifications of Nitko (1970) used as guidelines: (a) "The classes of behaviors that define different achievement levels are specified as clearly as is possible before the test is constructed [p. 38]." Behaviors defining the different logical developmental behavior levels

within each scale were worked out according to data provided by Inhelder and Piaget (1958), where each scale corresponded to a specific Piagetian task: exclusion - pendulum; proportion - balance; combination - colored and colorless liquids. According to Inhelder and Piaget, children manifest different logical behaviors on each task, depending upon the period of cognitive development that they are in. For each scale, each subscale corresponded to the cognitive behaviors characteristic of a different subperiod for the developmental logic of that scale. [See Gray (1970) for a summary of the three scales, Piagetian (sub) periods, and developmental logic used on the test.] (b) ". . . each behavior class is defined by a set of test situations (that is, test items or test tasks) in which the behaviors can be displayed in terms of all their important nuances [p.38]." For each logical scale, each developmental level (subscale) was defined by all of the written items that had the same logical structure as those logical behaviors characteristic of the specific Piagetian subperiods for the corresponding Piagetian tasks. Although the logical structure of the items was the same, the content was different. (c) ". . . given that the classes of behavior have been specified and that the test situations have been defined, a representative sampling plan is designed and used to select the test tasks that will appear on any form of the test [p. 38]." A total of thirty-six items were selected and adapted from those used in a pilot study. Each item had five alternatives, with the fifth alternative (e) always being "None of the above answers is correct."

Distribution of correct alternatives were as follows: A, B, D = 8 each, C = 7, E = 5. Items were randomly assigned their item number. Twelve items corresponded to the developmental logic of the pendulum, 12 items corresponded to the developmental logic of the balance, and 12 items corresponded to the developmental logic of combinations of colored and colorless liquids. Each scale of 12 items was divided into 6 items reflecting the logic of concrete operations (3 items for beginning concrete - concrete I; 3 items for concrete - concrete II) and 6 items reflecting the logic of formal operations (3 items for beginning formal - formal I; 3 items for formal - formal II). (d) ". . . the obtained score must be capable of expressing objectively and meaningfully the individual's performance characteristics in these classes of behavior [p. 38]." For each scale, subjects were given scores based on their patterns of correct and incorrect responses. For example, a subject classified as concrete II on the logic of combinations could use the logic of one-to-many and one-to-one logical multiplication and generally could not use the logic of combinations or permutation.

The general test directions and each item were controlled for reading difficulty by applying the Dale-Chall Readability Formula, with all but three of the items rated as fourth grade or lower. The remaining three items were rated at fifth - sixth grade difficulty.

Sample

Subjects were stratified by age by rounding their ages off to the nearest whole age. Within each age level a random sam-

ple of twelve subjects per age was selected, for a total of 96 subjects from 9 - 16 years of age. No student who was known to have reading problems was included.

Procedures

For each age level, one-half of the subjects were given the following Piagetian tasks: (a) Oscillation of a Pendulum, (b) Equilibrium in the Balance, and (c) Combinations of Colored and Colorless Chemical Bodies, first; and the written test, second. The remaining subjects were given the written test first and the Piagetian tasks second. Administration of the Piagetian tasks followed the guidelines "suggested" by Inhelder and Piaget (1958). All verbalizations were audio recorded, and the experimenter rated each subject's competence on each task on a behavioral rating sheet designed in accordance with the developmental level characteristics of subjects working with the three problems (Inhelder & Piaget, 1958). After one-half of the subjects in each age group had been tested with the Piagetian tasks, the written test was given to all subjects in a large group situation.

On each Piagetian task and each scale, subjects were classified as preoperational, concrete I, concrete II, formal I, or formal II. Classification criteria for the Piagetian tasks were those used by Inhelder and Piaget (1958). Classification criteria for each written scale were adapted from Longeot's (n. d.) and based on subscale-scale response patterns: preoperational - less than two correct responses for each subscale; concrete I - at least two correct responses on the

concrete I items and less than two correct responses on each of the other subscales; concrete II - at least two correct responses on each set of concrete items and less than two correct responses on each set of formal items; formal I - at least two correct responses on each set of concrete items and the formal I items, and less than two correct responses on the formal II items; formal II - at least two correct responses on each set of items.

The criteria were not met by 13.5% (39/288) of the patterns. Of the 39 non-ideal patterns, 18 were easily classifiable, leaving only 7.29% (21/288) response patterns which did not meet the classification criteria and were considered to be difficult patterns to classify.

Results

For each type of logic, there was no significant transfer from one type of test (Piagetian, written) to the other. Subjects who took the Piagetian tasks first and the written test second did no better than subjects taking the written test first and the Piagetian tasks second (Exclusion - Pendulum, $t = 1.02$; Proportion - Balance, $t = -.13$; Combination - Chemicals, $t = .81$, $df = 94$).

Convergent and Discriminant Validity

A multitrait-multimethod matrix (Campbell & Fiske, 1959) for the intercorrelations among the three Piagetian tasks and the three scales on the written test appears in Table 1. All

---

Insert Table 1 About Here

---

off-diagonal entries are significant ($p < .005$, $df = 94$, one-tail). Estimates of KR_{20} reliabilities are moderate, except for the written combinations. The validity values are of moderate size and greater than those considered substantial by Campbell & Fiske (1959), consequently Campbell and Fiske's criterion for convergent validity was met.

Evidence of the uniqueness of each set of logical behaviors from the others is less clear. All of the validity values meet Campbell and Fiske's first two criteria for discriminant validity--validity values should be greater than the respective row-column entries in the heteromethod triangles and the monomethod triangles--but not for all comparisons (See Table 2.).

Insert Table 2 About Here

All of the comparisons that did not meet the criteria involved a measure of the logic of exclusion; and the differences between the entries was small, the greatest having an absolute value of .022 (proportion validity vs. proportion-exclusion value in written monomethod triangle). The pattern of intercorrelations within the respective triangles also is not clear, as the patterns in the heteromethod triangles are different from each other and also different from the patterns in the monomethod triangles, which are the same.

For each set of developmental logic, there is definite evidence of convergent validity, but little evidence of discrimin-

ant validity, even though Campbell and Fiske (1959) state that the second of their discriminant validity criteria--validity values should be greater than respective row-column entries in the monomethod triangles---is an ideal criterion and not generally met.

Written Test

Table 3 presents mean item rankings for the three scales. For each scale, Pearson r's between the mean predicted rank

---

Insert Table 3 About Here

---

for each item and the mean empirical rank for each item were computed. All three correlations are significant, but only in the combination scale is there no interchanging of items from the different subscales. The two items from adjacent subscales with a difficulty of 16 are the only possible exceptions. Note that 8 of the combination items are extremely difficult, whereas only one item from the other scales is as difficult. The "cellar effect" definitely restricted the range of scores for the written combination scale, resulting in the medium low correlations and reliability for the scale (See Table 1.). In effect, the correlations involving the written combination scale were artificially depressed and, in reality, are probably much higher.

Age and Sex

For each of the written scales and each Piagetian task, a one-way ANOVA with unequal cell frequencies was run across ages,

with an ANOVA performed for each sex and the total sample. Scheffé's technique was then applied to the data of those ANOVA's that were significant at .05. The ages between which the greatest increase in mean classifications occurred was determined. Age levels below the increase were considered as one comparison group, while age levels above the increase were considered as the second comparison group. Comparison groups were chosen with the assumption that the ages between which the greatest increase in mean classification occurred reflected the ages at which the majority of subjects made the transition from concrete operations to formal operations. Consequently, the Scheffé comparisons were assumed to be between concrete operational and formal operational subjects. Table 4 summarizes these results.

Insert Table 4 About Here

In all cases where the original ANOVA was significant, the Scheffé comparison was significant at least at the level ($p < .10$) suggested by Scheffé (Ferguson, 1971) and, in most cases, at a lower level of probability. The ages at which the greatest increase in scores occurs is generally different, depending on the developmental logic measured, the method of assessment, and the sex of the subjects, although the greatest number of "jumps" in mean scores occurred between twelve and fourteen years of age.

Discussion

The correlations between the two methods measuring the same

set of developmental logic (validity values) along with moderate reliabilities are encouraging in that they are sufficiently large to support the conclusion that a written test using the developmental logic postulated by Piaget as its behavioral criterion is definitely possible, although there is room for improvement in this particular attempt. Also, the evidence of convergent validity is supportive of the generalization of Piagetian theory to "non-Piagetian" tasks. This lends credence to Piaget's belief that his conception of developmental levels is evidenced in Piagetian tasks and other tasks (Inhelder & Piaget, 1958). Psychometrically, the lack of discriminant validity of the developmental logics is disappointing and would indicate a definite effect of method variance (See Table 1.); yet this same lack of clearcut discrimination between the different sets of developmental logic provides support for Piaget's contention that a set of developmental logic is only one manifestation of an individual's general reasoning level; and generally when one set of logic has developed, other logics characteristic of that period should also have developed [See logic in Inhelder and Piaget (1958) and Gray (1970).].

The correspondence between the predicted and empirical written item sequences is excellent, indicating that Piaget's hypothesis of the developmental sequence of logical behaviors can be measured using Piagetian-based logical written problems. An exception is the exclusion concrete II items, on the average, being easier than the concrete I items. Both sets of items were

serializations, with the concrete I items composed of three entities for comparison and the concrete II items composed of four entities for comparison. The three items using the logical comparison "greater than' irrespective of number of entities to be compared were all easier than the three items using the logical comparison "less than";but a Scheffé comparison between the mean difficulties of the two sets was not significant ($\Phi/\sigma_{\Phi} = 1.49$, $p < .10$).

The extremely low difficulties and validity value for the logic of combinations would seem to indicate that the recognition-oriented multiple-choice format is not sensitive enough to measure the combinatorial ability of subjects. Rather, it appears, based on work by Longeot (1962, 1964, n. d.) and current work of the author, that an open-ended type of question, where the subject is required to generate the combinations, is much more sensitive in measuring combinatorial ability. The open-ended type question is certainly more "in the spirit" of Piaget, where the subject generally has to generate his own answers and not select the "best one" from a predetermined list.

Evidence of the subjects' possible past experience with written proportion types of questions can be seen in the proportion "cutoff" ages in Table 4. The "cutoff" age for the written proportions across sex and total sample is consistently a minimum of two years younger compared to the "cutoff" ages for its logical counterpart--the balance--and any other comparison with the exception of males and total sample for the pendulum. This would indicate that the written proportions may be tapping past

specific learnings as well as the logical operations of proportions, although such staggering of the "cutoff" ages is not uncommon (See Lovell, 1971; Piaget, 1971.) or unreasonable, and the "cutoffs" reported are generally consistent with previously reported results for similar type items (Winch, 1922a, 1922b; Burt, 1919; Longeot, 1962, 1964).

It appears that a Piagetian-based written test of logical cognitive development is possible if it is constructed according to behaviorally-oriented guidelines for criterion-referenced measurement. Certainly such a test is desireable, considering the traditional problems of evaluating cognitive skills and the problems associated with adequate measurement of skills in such individualized instruction programs as IGE. If such a test can be refined, a series of developmentally-based criterion-referenced tests which would demand the same cognitive skills, but for different content areas, could be constructed. Such a series of tests would have an advantage over current tests of being able to more accurately determine the reasoning level of a student within a specific content domain, and, hopefully, facilitate instruction and learning. At worst, it would be a device based on the <u>actual</u> cognitive development of children rather than something that is merely statistically convenient.

REFERENCES

Burt, C. The development of reasoning in school children--I. Journal of Experimental Pedagogy and Training College Record, 1919, 5, 68-77.

Campbell, D. T. & Fiske, D. W. Convergent and discriminant validation by the multitrait-multimethod matrix. Psychological Bulletin, 1959, 56, 81-105.

Ferguson, G. A. Statistical analysis in psychology & education. (3rd ed.) New York: McGraw-Hill, 1971.

Glaser, R. & Cox, R. C. Criterion-referenced testing for the measurement of educational outcomes. In R. Weisgerber (Ed.), Instructional process and media innovation. Chicago: Rand McNally, 1968. Pp. 545-550.

Glaser, R. & Klaus, D. J. Proficiency measurement: Assessing human performance. In R. Gagné (Ed.), Psychological principles in systems development. New York: Holt, Rinehart, and Winston, 1962, Pp. 419-474.

Gray, W. M. Children's performance on logically equivalent Piagetian tasks and written tasks. Educational Research Monographs. Dayton, Ohio: University of Dayton, 1970.

Inhelder, B. & Piaget, J. The growth of logical thinking from childhood to adolescence An essay on the construction of formal operational structures. New York: Basic Books, 1958.

Longeot, F. An essay of application of genetic psychology to differential psychology. (Translated by K. Kelley from Un essai d'application de la psychologie génétique à la psychologie différentielle. B.I.N.O.P., 1962, 18, 153-162.

Longeot, F. Statistical analysis of three collective genetic tests. (Translated by K. Kelley from Analyse statistique de trois tests génétiques collectifs. B.I.N.O.P., 1964, 20, 219-232.

Longeot, F. Test of Anagrams; Formal Combinational Operations; Formal Operations-Logic of Propositions; Formal Operations-Probabilities. (Translated by K. Kelley) n.d.

Lovell, K. Intellectual growth and understanding mathematics. Paper presented at the annual meeting of the American Educational Research Association, New York, February 1971.

Nitko, A. J. Criterion-referenced testing in the context of instruction. In Testing in turmoil: A conference on problems and issues in educational measurement. The thirty-fifth annual conference of the Educational Records Bureau. Greenwich, Connecticut: Educational Records Bureau, 1970. Pp. 37-40.

Piaget, J. The theory of stages in cognitive development. In D. R. Green, M. P. Ford, & G. B. Flamer (Eds.), Measurement and Piaget Proceedings of the CTB/McGraw-Hill conference on ordinal scales of cognitive development. New York: McGraw-Hill, 1971. Pp. 1-11.

Winch, W. H. Children's reasonings: Experimental studies of reasoning in school-children. Part II. Journal of Experimental pedagogy and Training College Record, 1922, 6, 199-212. (a)

Winch, W. H. Children's reasonings: Experimental studies of reasoning in school-children. Part III. Journal of Experimental Pedagogy and Training College Record, 1922, 6, 275-287. (b)

Footnotes

[1]The research reported is based upon a dissertation submitted to the State University of New York at Albany and partially supported by the University of Dayton Research Council Grant No. 94060-60. The author expresses his appreciation to Dr. Richard M. Clark for his guidance in the execution of this project.

TABLE 1

Multitrait-Multimethod Matrix*

		Piagetian			Written		
		E	P	C	E	P	C
Piagetian	E						
	P	.612					
	C	.531	.492				
Written	E	.661	.510	.519	(.785)		
	P	.602	.593	.452	.615	(.791)	
	C	.351	.384	.517	.443	.439	(.436)

Note.--Entries within parentheses are estimates of KR_{20}. Total test reliability estimate of KR_{20} = .867. Heterotrait-monomethod triangles are enclosed by solid lines, heterotrait-heteromethod triangles are enclosed by broken lines. E = exclusion, P = proportion, C = combination.

* All off-diagonal entries are significant ($p < .005$, df = 94, one-tail).

TABLE 2

Validity Values Meeting Campbell & Fiske's

First Two Criteria for Discriminant Validity

Logical Structure	Heterotrait-Heteromethod	Heterotrait-Monomethod
Exclusion	4/4	4/4
Proportion	3/4	2/4
Combination	3/4	3/4

TABLE 3

Predicted and Empirical Sequences of Written Items

													r
Exclusion													
Predicted	2	2	2	5	5	5	3	8	8	11	11	11	.876*
Empirical	4.33 (70)	4.33 (52)	4.33 (45)	2.67 (75)	2.67 (59)	2.67 (56)	8 (41)	8 (36)	8 (36)	11 (31)	11 (27)	11 (24)	
Proportion													
Predicted	2	2	2	5	5	5	8	8	8	11	11	11	.976*
Empirical	3 (85)	3 (74)	3. (69)	4 (74)	4 (74)	4 (68)	8.33 (49)	8.33 (44)	8.33 (39)	10.67 (43)	10.67 (26)	10.67 (6)	
Combination													
Predicted	2	2	2	5	5	5	8	8	8	11	11	11	.999*
Empirical	2 (94)	2 (91)	2 (88)	5.17 (48)	5.17 (17)	5.17 (16)	7.83 (16)	7.83 (15)	7.83 (11)	11 (4)	11 (3)	11 (3)	

Note.--Main entries are the mean ranks for the item's subscale. Values in parentheses are the difficulties of each item.

* $p < .005$, one-tail, df = 10

TABLE 4

Summary of Scheffé Comparisons

	Male df = 7, 46	Female df = 7, 34	Total df = 7, 88
Exclusion	4.70^{4} (9 - 13) vs. (14 - 16)	4.16^{2} (9 - 12) vs. (13 - 16)	6.24^{6} (9 - 13 vs. (14 - 16)
Proportion	5.12^{5} (9 - 11) vs. (12 - 16)	5.17^{5} (9 - 10) vs. (11 - 16)	6.91^{6} (9 - 11) vs. (12 - 16)
Combination	ANOVA n. s.	ANOVA n. s.	4.26^{3} (9 - 13) vs. (14 - 16)
Exclusion (Pendulum)	5.54^{5} (9 - 12) vs. (13 - 16)	6.83^{6} (9 - 12) vs. (13 - 16)	8.69^{6} (9 - 12) vs. (13 - 16)
Proportion	5.13^{5} (9 - 13) vs. (14 - 16)	3.76^{1} (9 - 12) vs. (13 - 16)	5.38^{6} (9 - 14) vs. (15 - 16)
Combination (Chemicals)	4.20^{2} (9 - 14) vs. (15 - 16)	ANOVA n. s.	5.27^{5} (9 - 14) vs. (15 - 16)

NOTE.--Intries are the absolute values of the Scheffé comparisons (Φ/θ_0). Entries in parentheses refer to age levels that were grouped for comparison. Degrees of freedom for each comparison are listed at the top of the respective columns.

$^{1}p < .10, ^{2}p < .05, ^{3}p < .025, ^{4}p < .01, ^{5}p < .005, ^{6}p < .001$